hESC-based human glial chimeric mice reveal glial differentiation defects in Huntington disease

Mikhail Osipovitch[1], Andrea Asenjo-Martinez[1], Adam Cornwell[2], Simrat Dhaliwal[2], Lisa Zou[2], Devin Chandler-Militello[2], Su Wang[2], Xiaojie Li[2], Sarah-Jehanne Benraiss[2], Robert Agate[2], Andrea Lampp[1], Abdellatif Benraiss[2], Martha S. Windrem[2], Steven A. Goldman[1,2,3*]

[1]Center for Translational Neuromedicine, University of Copenhagen Faculty of Health and Medical Science, 2200 Copenhagen N, Denmark; [2]Center for Translational Neuromedicine, University of Rochester Medical Center, Rochester, NY, 10021, USA; [3]Neuroscience Center, Rigshospitalet-Copenhagen University Hospital, Copenhagen, Denmark.

Materials and Methods for HD RNA-Seq Analysis

Embryonic stem cells (hESCs) derived from 3 Huntington's disease embryos (designated to HD lines 17, 18, and 20) and 2 healthy control embryos (designated to CTR lines 02 and 19) were obtained from Genea Biocells, Sydney, Australia (http://geneabiocells.com/services/shelf-products/human-embryonic-stem-cells/). The cell lines CTR02 and HD20 comprised a sibling pair. The hESC lines were differentiated into glia by previously described methods (Wang et al., 2013) and further purified by FACS targeting CD140a for enriched populations of Glial Progenitor Cells (GPCs) and CD44 for enriched populations of Astrocyte Progenitor Cells (APCs). The purified glial cell populations where then used in mRNA isolation by PolyA selection and mRNA sequencing analysis. Sequencing libraries were prepared using TruSeq RNA v2 kit. The libraries were sequenced on Illumina HiSeq 2500 platform for approximately 45 million of 100-bp single-end reads per sample (GPC lines CTR19, HD17, HD18, and HD20) and for similar depth but of 2x125-bp paired-end reads (GPC line CTR02 and all APC lines). The sequencing reads were then pre-processed by trimming off adapter and low-quality sequences using Trimmomatic [3]. The quality of reads before and after pre-processing was assessed with FastQC [1]. The pre-processed reads were then aligned to the RefSeq NCBI reference [10] human genome version GRCh38 with Subread read aligner [7]. Raw gene counts were obtained from BAM alignment files with featureCounts [8].

After examining principal component and hierarchical clustering plots generated with native R functions [11] 1 mis-clustered outlier sample was removed from further analysis in GPC line HD17 and 2 outlier samples in GPC lines HD20 and CTR19 each. After eliminating lowly expressed transcripts leaving those with a count of at least 5 reads in more than 3 samples, the count data were normalized using RUVSeq [12] R Bioconductor [4] package to account for variance. As described in RUVSeq manual, the normalization was accomplished in the following three-step procedure: *(1)* negative *in silico* control genes were determined by first-pass differential expression analysis by edgeR [13] and DESeq2 [9] R Bioconductor packages taking genes with FDR-adjusted P Values of above 0.75 as calculated by both methods; *(2)* the negative control genes were then used in RUVg function of RUVSeq package for calculation of variance factors; and *(3)* the second-pass differential expression analysis (1% FDR and log2 fold change > 1) for determination of disease-dysregulated genes was performed using the original raw counts with adjusting for RUVg-calculated variance factors by multi-factor GLM models implemented in edgeR and DESeq2 packages.

The filtering for lowly expressed transcripts and the three-step analysis procedure were employed for comparisons of each HD-derived cell line to pooled CTR-derived line and for the sibling pair comparison of HD20 vs. HD19, within each of the two cell types. In all comparisons, 1 RUVg-calculated variance factor was used. Within each cell type, the intersection of the resulting four lists of differentially expressed genes was taken as the conserved representative list of HD-dysregulated genes. For all differential expression

comparisons, only the significant results that agreed between edgeR and DESeq2 methods were used in downstream analysis. Fold changes and FDR-adjusted P Values reported in the results section were calculated by edgeR. Functional annotation of the conserved set of HD-dysregulated genes was performed in ToppCluster [6] and Ingenuity Pathway Analysis (IPA) software [5]. Network visualization and analysis of the results of functional annotation were performed in Gephi [2] graph visualization software. The complete reproducible workflow, including R scripts and count matrix, was deposited to https://github.com/cbtncph/HD-Glial-Differentiation-Block-Goldman-Lab-2017.

Citations in Materials and Methods for HD RNA-Seq Analysis

[1] Andrews S. (2010). FastQC: A quality control tool for high throughput sequence data. *Reference Source*.

[2] Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM 8*, 361-362.

[3] Bolger A.M., Lohse M., & Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

[4] Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., ... & Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), R80.

[5] Ingenuity Pathway Analysis, QIAGEN Redwood City, www.qiagen.com/ingenuity

[6] Kaimal V. et al. (2010). ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Research*, gkq418.

[7] Liao Y., Smyth G.K., Shi W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108.

[8] Liao Y., Smyth G.K., Shi W. (2014). featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923-30.

[9] Love M.I., Huber H., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome biology*, 15(12), 1-21.

[10] Pruitt K.D., Tatusova T., & Maglott D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1), D61-D65.

[11] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[12] Risso D., Ngai J., Speed T., Dudoit S. (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples." *Nature Biotechnology*, 32(9), pp. 896–902.

[13] Robinson M.D., McCarthy D.J., & Smyth G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139-140.